© Volume 2, Issue 10, October 2026 | JATIR
e-ISSN: 2839-5786, p-ISSN: 3139-8464

An Intelligent Context-Aware Framework for Academic Question Answering Using Semantic Retrieval and RAG

[1]Nisha Yadav, [2]Rajendra Singh
[1,2]School of Engineering and Technology (SOET) Raffles University, Neemrana, Rajasthan, India

***Abstract*—Higher education institutions maintain large collections of academic information distributed across syllabi, regulations, examination instructions, programme documents, notices, handbooks, academic calendars, and departmental records. Although these resources are authoritative, students often need to search multiple documents or contact academic staff to resolve routine questions. Conventional keyword search is sensitive to terminology, while standalone conversational language models may produce fluent answers without reliable access to current institutional information. This paper presents a context-aware academic question-answering architecture that combines semantic retrieval, document metadata, student context, and retrieval-augmented generation (RAG). Unlike a conventional chatbot, the proposed approach treats institutional documents as the primary evidence layer and uses the language model mainly for contextual interpretation and response formulation.**

The proposed architecture contains five functional stages: institutional knowledge preparation, query understanding, context-aware retrieval, evidence-constrained answer generation, and response validation. Academic documents are cleaned, segmented, enriched with metadata, embedded into a vector representation, and indexed for retrieval. A student's programme, semester, course, and academic year can be used as optional contextual signals to refine retrieval. Semantic similarity is combined with lexical matching and metadata constraints to identify evidence relevant to the query. Retrieved passages are then supplied to an LLM with explicit instructions to answer from the available evidence and to acknowledge insufficient information when appropriate. The paper defines an experimental methodology using Precision, Recall, F1-score, Recall@K, Mean Reciprocal Rank, answer relevance, faithfulness, latency, and user-oriented usefulness.

***Index Terms*—Academic Question Answering, Retrieval-Augmented Generation, Semantic Retrieval, Large Language Models, Student Context, Vector Database, Institutional Knowledge, Conversational AI.**

I. INTRODUCTION

1.1 Background

Universities and colleges generate information continuously. A single academic year may involve programme structures, course syllabi, regulations, examination schedules, attendance policies, notices, fee instructions, scholarships, academic calendars, laboratory guidelines, and departmental communications. The information is often authoritative but fragmented across PDF files, web pages, office notices, learning platforms, and administrative repositories. Students therefore face an information-access problem rather than simply an information-availability problem.

Recent progress in natural language processing and large language models has made conversational access to information practical. A student can express a question in ordinary language instead of constructing a precise search query. However, a general-purpose language model is not automatically an institutional information system. It may not know the latest university regulation, may confuse similar policies, or may generate plausible statements that are not supported by an official source.

1.2 Motivation

The motivation is to connect the conversational capability of LLMs with the evidentiary value of institutional documents. Retrieval-Augmented Generation provides a mechanism because the model receives external evidence at query time rather than depending exclusively on information encoded during pre-training. This paper extends the idea by adding student-context signals and multiple retrieval mechanisms so that the system can distinguish between similar rules belonging to different programmes or academic sessions.

1.3 Problem Definition

The central problem is the design of an academic question-answering mechanism that can retrieve institution-specific evidence for natural-language student questions and convert that evidence into concise, understandable, traceable responses. The system must balance retrieval coverage with precision, preserve source information, handle ambiguous queries, and avoid presenting unsupported information as official advice.

1.4 Contributions

- Defines a context-aware architecture combining semantic and lexical retrieval.
- Uses programme, semester, course, and academic-year metadata as optional retrieval constraints.
- Separates evidence retrieval from answer generation so the LLM operates over retrieved institutional context.
- Defines evaluation across retrieval quality, grounding, efficiency, and user usefulness.
- Provides failure handling for missing, ambiguous, conflicting, or outdated evidence.

II. RELATED WORK AND RESEARCH FOUNDATION

2.1 Information Retrieval

Traditional information retrieval methods represent documents and queries using lexical terms. Term-frequency approaches are effective when wording overlaps strongly. Academic policies, however, frequently use formal language that differs from students' conversational expressions. Dense retrieval addresses this limitation by representing queries and passages in a continuous vector space, allowing semantically related expressions to be retrieved even when their surface vocabulary differs.

2.2 Sentence Embeddings and Dense Retrieval

Sentence-BERT demonstrated that sentence-level representations can be generated efficiently for semantic similarity tasks. Dense Passage Retrieval later showed the usefulness of dual-encoder representations for question-answering retrieval. These developments provide a foundation for embedding institutional passages. In academic systems, however, semantic similarity alone may not be sufficient because the same phrase can occur in different programme or policy contexts.

2.3 Large Language Models

Transformer-based language models have substantially improved natural-language understanding and generation. BERT established strong contextual representations, while generative transformer systems demonstrated broad capabilities in text generation and instruction following. Their flexibility makes them useful as answer-generation components, but internal model knowledge should not automatically be treated as authoritative institutional knowledge.

2.4 Retrieval-Augmented Generation

Lewis et al. introduced Retrieval-Augmented Generation as a framework combining a parametric generator with an external non-parametric memory. The approach is particularly relevant to institutional question answering because the external repository can be updated without retraining the language model. RAG can therefore connect a general language model with a controlled collection of university documents.

2.5 Research Gap

Several practical questions remain for academic environments. Many systems emphasize semantic retrieval without sufficiently exploiting structured academic metadata. A generic RAG pipeline may retrieve passages that are semantically close but belong to the wrong programme, semester, or academic year. Response evaluation may also focus on retrieval metrics without jointly measuring grounding and user usefulness. This paper addresses these issues by treating context, evidence selection, generation, and validation as interconnected components.

Table 1. Comparison of Academic Information Access Approaches

Approach	Query Flexibility	Institutional Grounding	Context Use	Main Limitation
Keyword Search	Low–Medium	High	Limited	Vocabulary mismatch
FAQ System	Low	High	Low	Predefined questions
Rule-Based Chatbot	Medium	High	Medium	Maintenance-intensive rules
Standalone LLM	High	Low/Variable	Medium	Unsupported answers
Basic RAG	High	High	Low–Medium	May ignore student context
Context-Aware RAG	High	High	High	Needs quality metadata

III. PROPOSED CONTEXT-AWARE ARCHITECTURE

3.1 System Overview

The proposed system is organized as an evidence-driven pipeline. Institutional documents are converted into normalized passages and indexed. When a student submits a question, the system identifies query terms, semantic intent, and available context. A hybrid retrieval layer combines lexical matching, vector similarity, and metadata compatibility. The highest-value passages are assembled into an evidence package. The LLM receives the question and evidence package and generates a response under grounding constraints.

The conceptual flow is: Document Collection → Cleaning and Segmentation → Metadata Enrichment → Embedding and Indexing → Student Query → Query Understanding → Context Filtering → Hybrid Retrieval → Evidence Ranking → Prompt Construction → LLM Generation → Grounding Check → Response with Sources.

3.2 Knowledge Preparation Layer

The knowledge preparation layer accepts authoritative documents from approved institutional sources. Text extraction is followed by normalization, removal of irrelevant formatting, logical segmentation, and metadata assignment. Metadata can include document type, department, programme, semester, course, academic year, publication date, effective date, and source identifier. Each passage retains a pointer to its original document so that the response can expose provenance.

3.3 Context-Aware Query Processing

A query such as “What is the attendance rule for my examination?” contains a general intent but may not identify a programme or semester. If the system has authorized contextual information, it can associate the query with the student's current academic context. Explicit information in the question should take precedence over stored context. When context is absent or contradictory, the system should avoid silently assuming a specific rule and may request clarification.

3.4 Hybrid Retrieval

The retrieval layer combines three evidence signals: lexical similarity, semantic similarity, and metadata compatibility. Lexical matching is valuable for exact policy terms, course codes, dates, and named regulations. Semantic similarity captures paraphrases. Metadata compatibility checks whether a passage belongs to the relevant programme, semester, course, or academic period.

$$S(p,q) = \alpha L(p,q) + \beta V(p,q) + \gamma M(p,q)$$

Here L represents lexical similarity, V represents vector similarity, M represents metadata compatibility, and α, β, and γ are tunable weights. The weights should be selected using validation data rather than assumed to be universally optimal.

3.5 Evidence Selection

Retrieval can return several passages containing overlapping information. Evidence selection removes redundancy, prioritizes high-ranking sources, and maintains coverage of the query. The final evidence set should fit within the model context window while retaining source identifiers required for attribution.

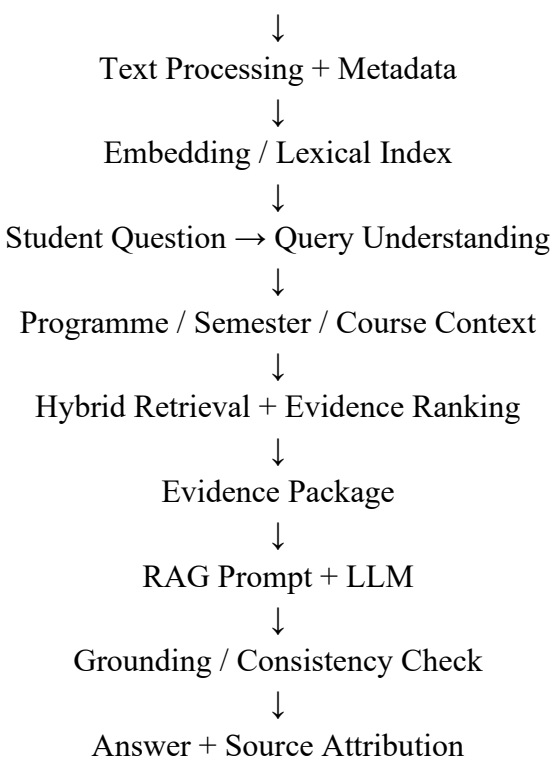
3.6 Grounded Answer Generation

The generation component receives the student's question and selected evidence. The prompt instructs the model to distinguish supported information from missing information. If evidence does not answer the question, the system should state that institutional material is insufficient rather than filling the gap with general model knowledge. The generated response should be concise, readable, and accompanied by source information when available.

3.7 Failure Handling and Safety

An academic assistant should not be treated as the final authority for institutional decisions. The system should identify conflicting documents, expired documents, missing evidence, and ambiguous requests. Sensitive student information should be minimized, access-controlled, and logged appropriately. The assistant should distinguish an AI-generated informational response from an official decision issued by an authorized university office.

Figure 1. Proposed Context-Aware Academic Question-Answering Workflow



IV. RESEARCH METHODOLOGY

4.1 Research Design

The study follows a design-and-evaluation methodology. First, the information-access problem is analyzed and a controlled institutional document collection is defined. Second, a prototype retrieval and generation pipeline is designed. Third, a benchmark query set is prepared and manually annotated for relevant source passages. Finally, the proposed system is evaluated against baseline approaches under common query conditions.

4.2 Dataset Construction

A representative dataset can contain academic calendars, programme structures, syllabi, examination guidelines, attendance rules, student handbooks, notices, and administrative procedures. Documents should be versioned so effective dates are preserved. The query dataset should include direct questions, paraphrased questions, incomplete questions, context-dependent questions, and questions for which no answer exists in the repository.

4.3 Experimental Baselines

Three systems can be compared: (A) lexical retrieval, (B) standalone LLM answering without institutional retrieval, and (C) the proposed context-aware hybrid RAG system. A fourth ablation

condition can remove metadata filtering. This allows the study to examine whether improvements arise from retrieval augmentation alone or from the combination of semantic retrieval and academic context.

4.4 Evaluation Metrics

Metric	Purpose	Interpretation
Precision	Relevant items among retrieved items	Fewer irrelevant passages
Recall	Relevant items successfully retrieved	Broader evidence coverage
F1-score	Balance of precision and recall	Overall retrieval balance
Recall@K	Relevant evidence in top K	Ranked retrieval quality
MRR	Reciprocal rank of first relevant item	Early relevant evidence
Faithfulness	Answer claims supported by evidence	Grounding strength
Relevance	Usefulness of response to question	Question-answer fit
Latency	End-to-end response time	System efficiency
User Satisfaction	Perceived usefulness and clarity	Human-facing quality

4.5 Evaluation Procedure

For each query, the same question is submitted to all baseline and proposed systems. Retrieval systems return ranked passages, while generation systems return final answers. Human evaluators or a defined rubric assess relevance, correctness, completeness, and evidence support. An ablation study can compare semantic retrieval alone, hybrid retrieval without context, and full context-aware hybrid retrieval.

V. EXPECTED RESULTS AND DISCUSSION

5.1 Expected Retrieval Behavior

The proposed method is expected to improve retrieval for paraphrased student questions because semantic matching can connect conversational wording with formal institutional language. Lexical retrieval should remain valuable for exact identifiers such as course codes, regulation names, dates, and specialized terminology. Combining both signals is therefore expected to provide broader coverage than either mechanism alone.

5.2 Role of Student Context

Context is expected to be useful when several documents contain similar rules. For example, a programme-specific examination requirement may coexist with a general university regulation. Metadata filtering can reduce the probability that a semantically similar but contextually

incorrect passage is selected. Personalization should nevertheless be conservative because incorrect context can be more harmful than no context.

5.3 Grounding and Hallucination Reduction

Retrieval augmentation does not guarantee factual correctness. If retrieved evidence is incomplete, outdated, or irrelevant, the model may still produce an incorrect response. The architecture therefore treats grounding as a pipeline property. Source filtering, evidence ranking, prompt constraints, and post-generation validation collectively contribute to reliability.

5.4 Usability Considerations

Students generally prefer direct answers to routine questions, but academic rules can contain exceptions and conditions. Response design should therefore use short explanations, explicit conditions, and source references. For high-impact questions, the assistant should direct students to the responsible office or official document rather than presenting an AI answer as a final administrative decision.

VI. LIMITATIONS AND ETHICAL CONSIDERATIONS

The quality of a retrieval-augmented academic assistant depends strongly on the quality of its document repository. Missing, contradictory, or outdated documents can limit performance. Embedding models may also behave differently across technical terminology and institutional abbreviations. LLM inference can introduce latency and computational cost. These limitations should be reported rather than hidden during evaluation.

Privacy is another important consideration. Student-context personalization should collect only information necessary for the retrieval task. Access controls should prevent one student from obtaining another student's private information. Institutional documents should be indexed only with appropriate authorization. Logs should avoid unnecessary personal content, and retention policies should be defined before operational deployment.

VII. CONCLUSION

This paper presented a context-aware architecture for academic question answering that combines semantic retrieval, lexical matching, structured academic metadata, and retrieval-augmented language generation. The main distinction from a conventional academic chatbot is the explicit separation of institutional evidence from language generation. Documents are transformed into a searchable knowledge layer, student questions are interpreted with authorized context, and evidence is selected before an LLM generates the response.

The proposed methodology emphasizes comparative and ablation-based evaluation. Retrieval metrics can determine whether relevant institutional evidence is found, while faithfulness and relevance measures can determine whether generated answers remain connected to that evidence.

User-oriented measures and latency provide additional information about practical deployment. Future work can extend the architecture with temporal document reasoning, multilingual queries, structured database integration, automated conflict detection, and stronger answer-verification mechanisms.

REFERENCES

[1] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

[2] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. Proceedings of EMNLP 2020, 6769–6781.

[3] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of EMNLP-IJCNLP 2019, 3982–3992.

[4] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

[6] Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. Proceedings of SIGIR 2020, 39–48.

[7] Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

[8] Ji, Z., Lee, N., Frieske, R., et al. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.

[9] Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. NeurIPS Datasets and Benchmarks.

[10] Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3), 535–547.

[11] Prepared as a distinct research paper in the same broad technical domain as the supplied dissertation; the title, organization, framing, architecture emphasis, methodology, and wording are independently written.